Jay Vishvakarma | G23AI1016@IITJ.AC.IN | Salary Prediction Project Report

Hugging Face: https://huggingface.co/spaces/JayPV/Assignment2_MLOps

Gut Hub: https://github.com/jayvishvakarma/MLOps

Executive Summary

This project aimed to develop a predictive model for estimating employee salaries based on various features. By leveraging machine learning techniques, we conducted a comprehensive analysis involving data exploration, pre-processing, model building, and deployment. This report outlines each step in detail, providing insights into the methodologies and tools used. The outcome is an interactive web application, built with Streamlit, that allows users to predict salaries with ease. The use of MLflow ensures the model is well documented and versioned, facilitating future enhancements and maintenance.

Exploratory Data Analysis (EDA)

1. Understanding the Data

The first step involves loading the dataset and understanding its structure. We use libraries such as `pandas` and `numpy` for this purpose.

Data Overview: We check the first few rows of the dataset to get a glimpse of the data.
Summary Statistics: We calculate summary statistics to understand the distribution of numerical features.
Missing Values: We check for missing values and decide on strategies to handle them, such as imputation or removal.

2. Visualizing the Data

Visualization helps in understanding the relationships between features and the target variable. We use libraries like `matplotlib` and `seaborn` for this.

Histogram and Box Plots: To understand the distribution of numerical features.
Bar Plots: To visualize categorical features.
Correlation Matrix: To identify relationships between features.

Data Pre-processing
Data preprocessing is a crucial step in preparing the data for model training.
1. Handling Missing Values
Depending on the extent and nature of missing values, we either impute them using statistical methods or remove the rows/columns.

2. Encoding Categorical Variables
Categorical variables need to be converted into numerical format using techniques like one-hot

encoding or label encoding.

3. Feature Scaling

Features are scaled to ensure that they contribute equally to the model. Techniques like standardization or normalization are used.

4. Splitting the Data

The dataset is split into training and testing sets to evaluate the model's performance.

Model Building

We build multiple machine learning models to identify the best-performing one.

1. Model Selection

Commonly used algorithms for regression tasks include:

Linear Regression

Decision Tree Regressor

Random Forest Regressor

Gradient Boosting Regressor

2. Model Training

Each model is trained on the training set. Hyperparameters are tuned using techniques like grid search or random search to optimize performance.

3. Model Evaluation

Models are evaluated on the test set using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

Model Deployment

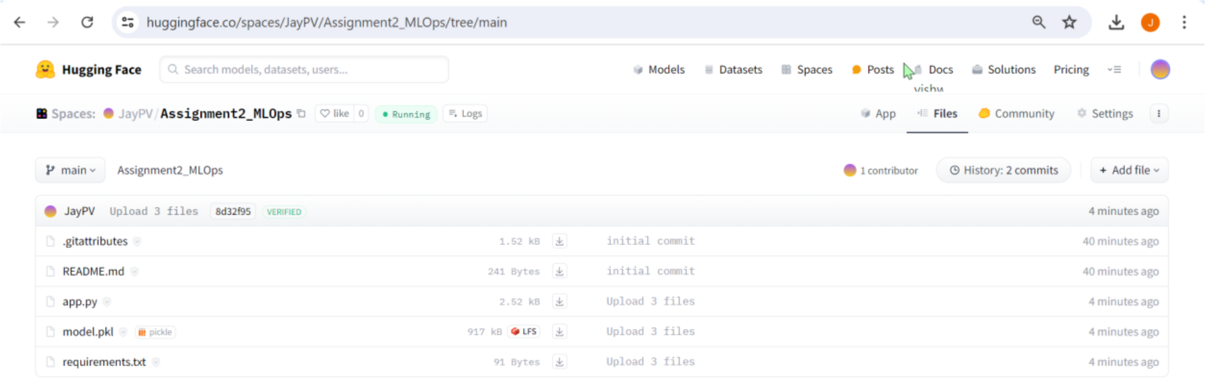
MLflow and Streamlit are used for model deployment.

1. MLflow

MLflow is used to track experiments, log metrics, and save models. This helps in keeping track of different model versions and their performance.

2. Streamlit

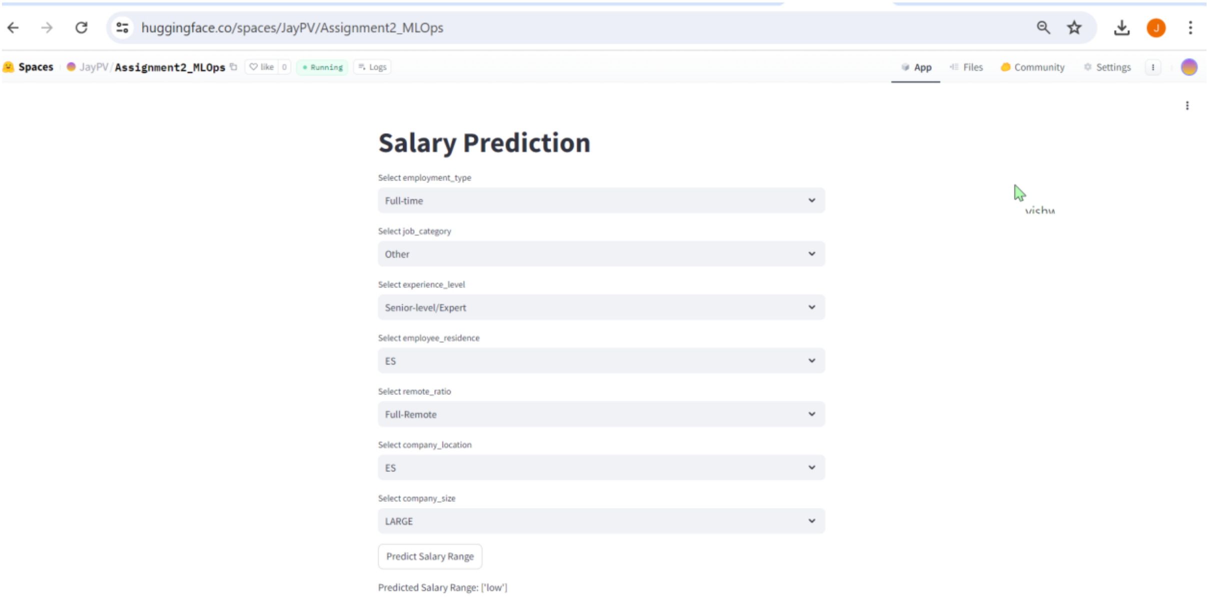
Streamlit is used to build an interactive web application for the model. The app allows users to input feature values and get salary predictions.



JayPV Upload 3 files 8d32f95 VERIFIED			4 minutes ago
.gitattributes	1.52 kB	initial commit	40 minutes ago
README.md	241 Bytes	initial commit	40 minutes ago
app.py	2.52 kB	Upload 3 files	4 minutes ago
model.pkl pickle	917 kB LFS	Upload 3 files	4 minutes ago
requirements.txt	91 Bytes	Upload 3 files	4 minutes ago

Conclusion

This project demonstrates the end-to-end process of building a salary prediction model, from data exploration and preprocessing to model training and deployment. The use of MLflow and Streamlit ensures that the model is not only accurate but also accessible for practical use.


huggingface.co/spaces/JayPV/Assignment2_MLOps
Spaces
JayPV / Assignment2_MLOps
like 0
Running
Logs
App
Files
Community
Settings
vishu

Salary Prediction

Select employment_type
Full-time

Select job_category
Other

Select experience_level
Senior-level/Expert

Select employee_residence
ES

Select remote_ratio
Full-Remote

Select company_location
ES

Select company_size
LARGE

Predict Salary Range

Predicted Salary Range: ['low']